Challenges of the LHC: Computing

Torre Wenaus
ATLAS Experiment / LCG Applications Area
BNL / CERN

The Defining Characteristic: Scale

Key drivers for LHC Computing
- Global enfranchisement in analysis
- Leveraging global resources
- Rapid analysis - don't hinder physics
- Global software development
- Effective communication, collaboration

LHC collaborating nations

Most computing resources not at CERN

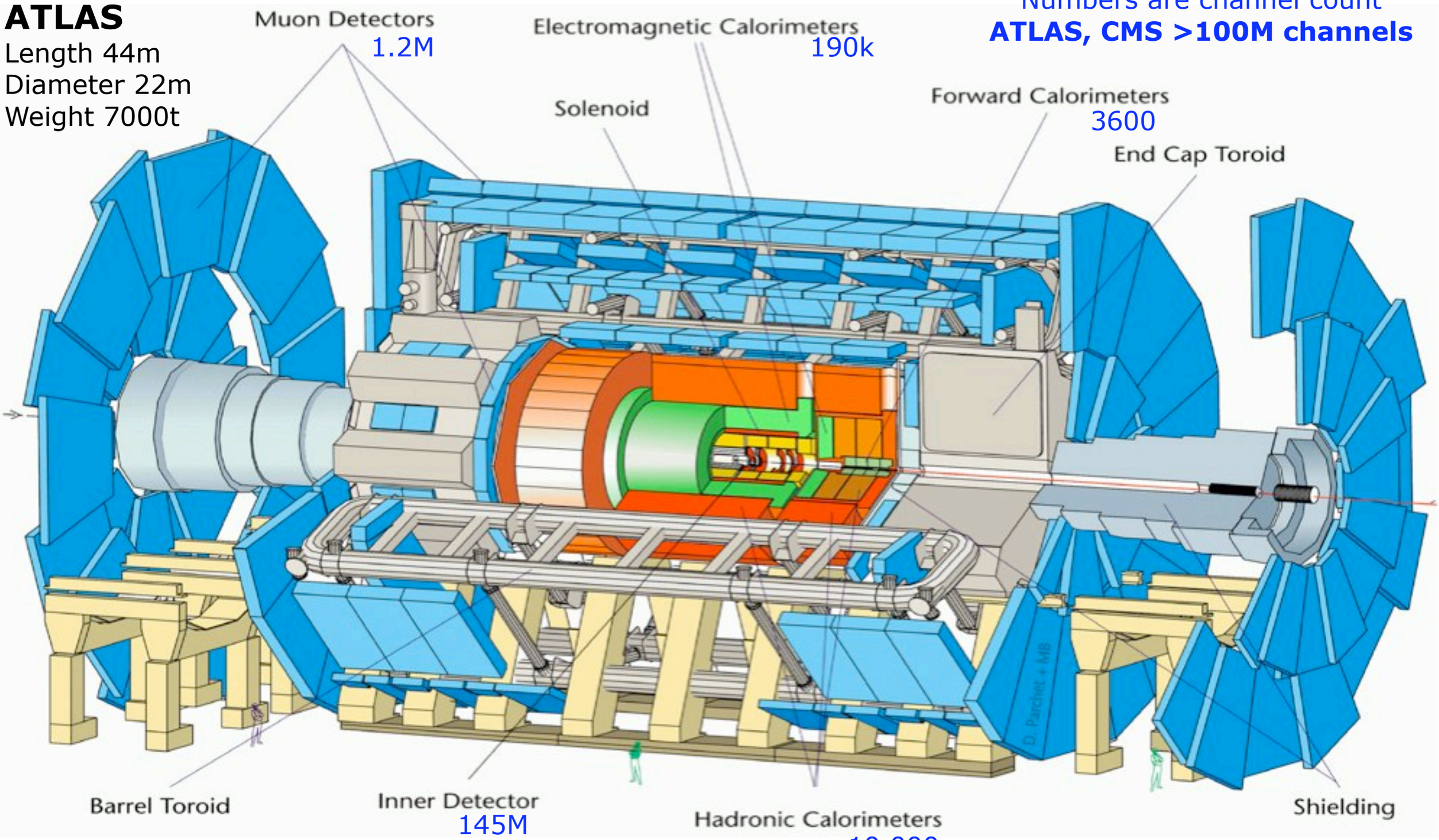
Some of ATLAS

~6000 LHC collaborators

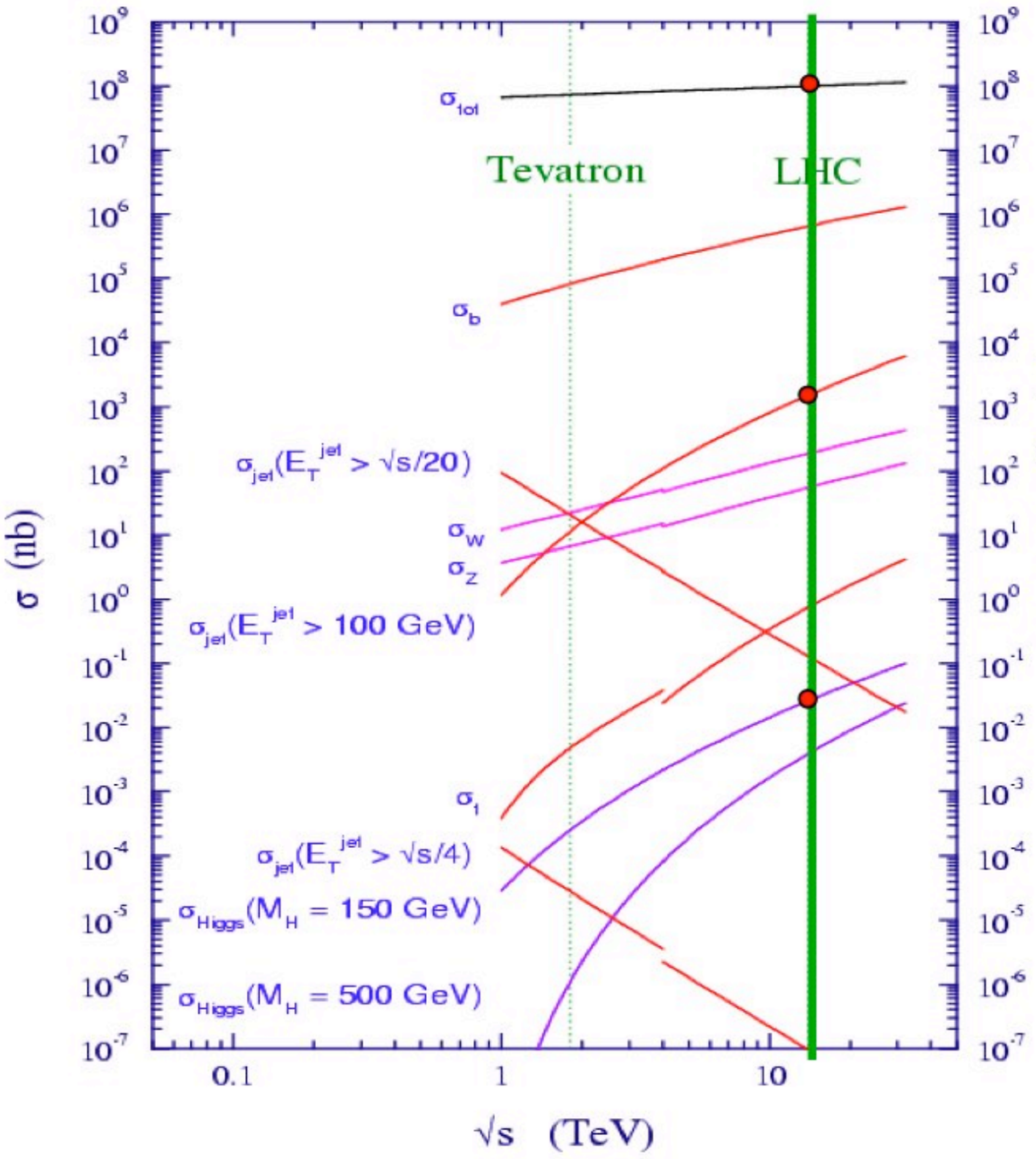
Detector Scale/Complexity

ATLAS
Length 44m
Diameter 22m
Weight 7000t

Numbers are channel count
ATLAS, CMS >100M channels



Low Rate Physics



Total cross section

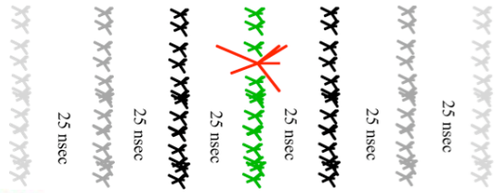
Key physics is low rate

Higgs, SUSY, ...

Demands
- high luminosity, data rates
- highly selective trigger
- offline sparse signal extraction

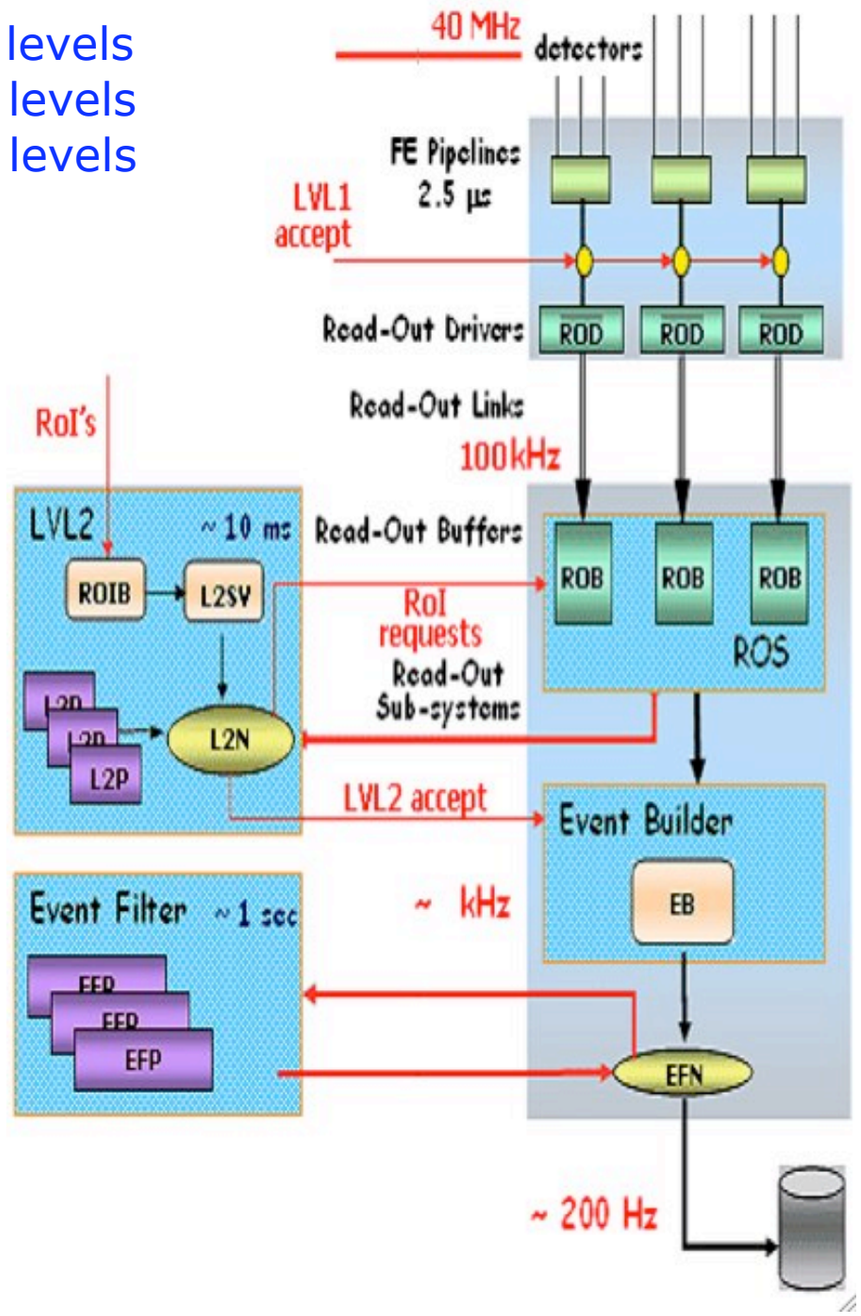
Rare signal extraction in huge data samples of complex detectors -- large scale processing and data management

Online Data Reduction - ATLAS



Bunch crossing every 25ns
Each with ~20 interactions

CMS: 2 levels
LHCb: 3 levels
ALICE:4 levels



40 MHz, 1 PB/sec

Level 1: Fast detector triggers
(calorimeter, muon trigger)

100 kHz, 100 GB/sec

Level 2: Full detector info
in (triggered) regions of interest

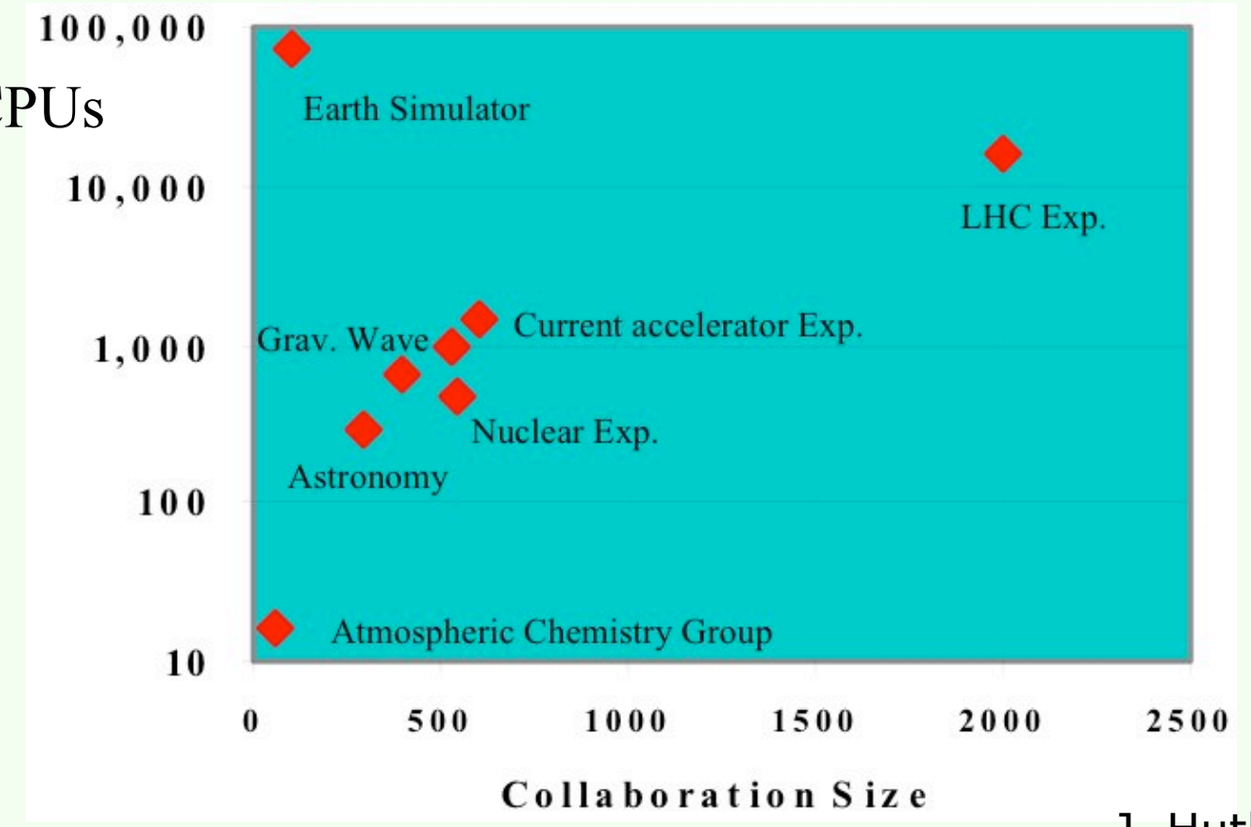
3 kHz, 4.5 GB/sec

Event filter: full event reconstruction
with calibration/alignment

~200 Hz, ~320 MB/sec recorded data

~3 PB/year raw data

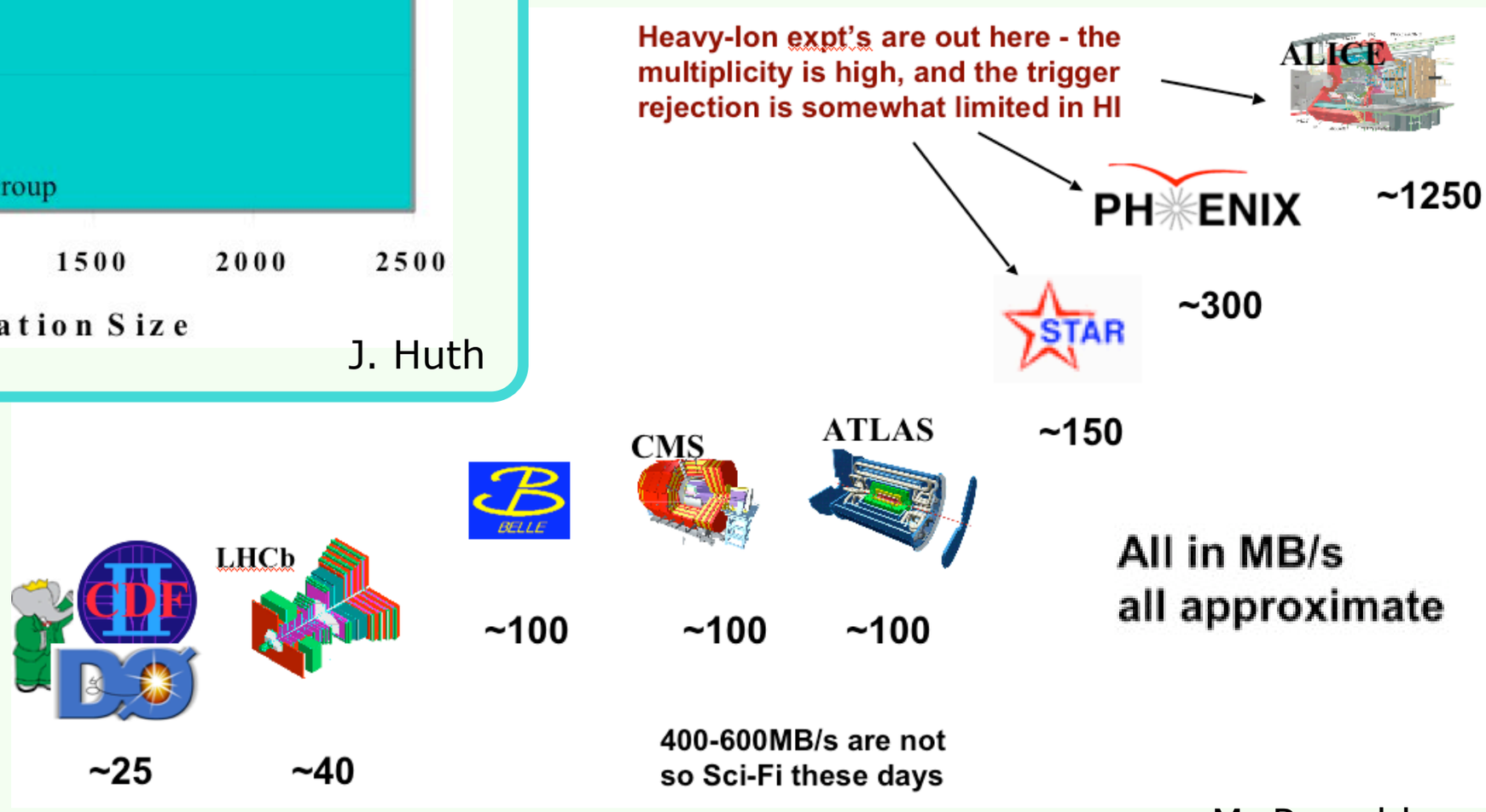
How LHC Computing Compares



J. Huth

Collaborators, Processing

Data Rates



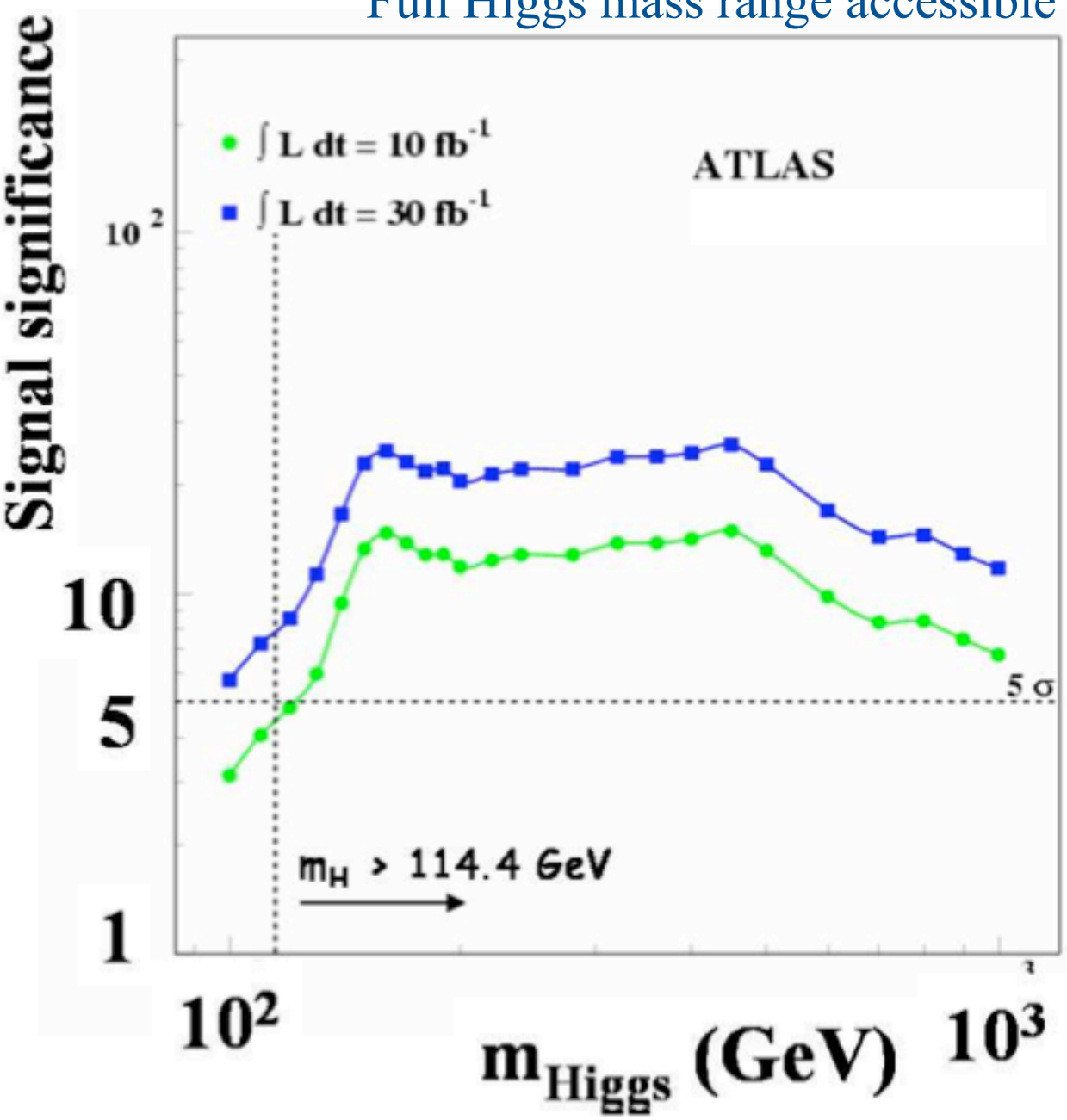
M. Purschke

Computing Must Be *Ready!*

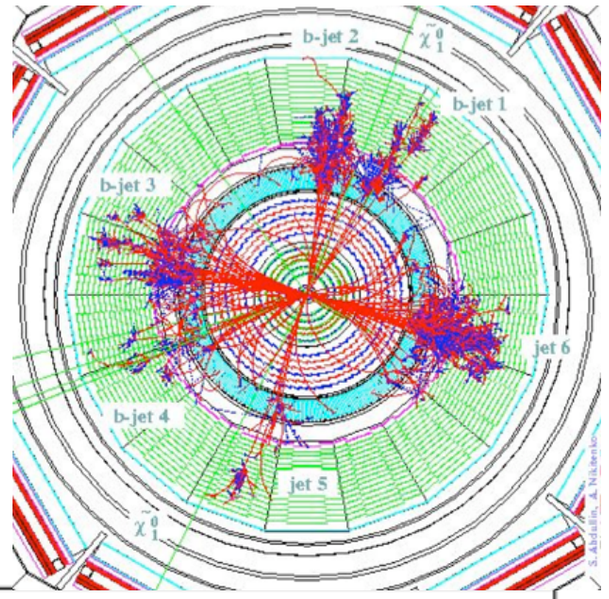
Tremendous discovery reach early

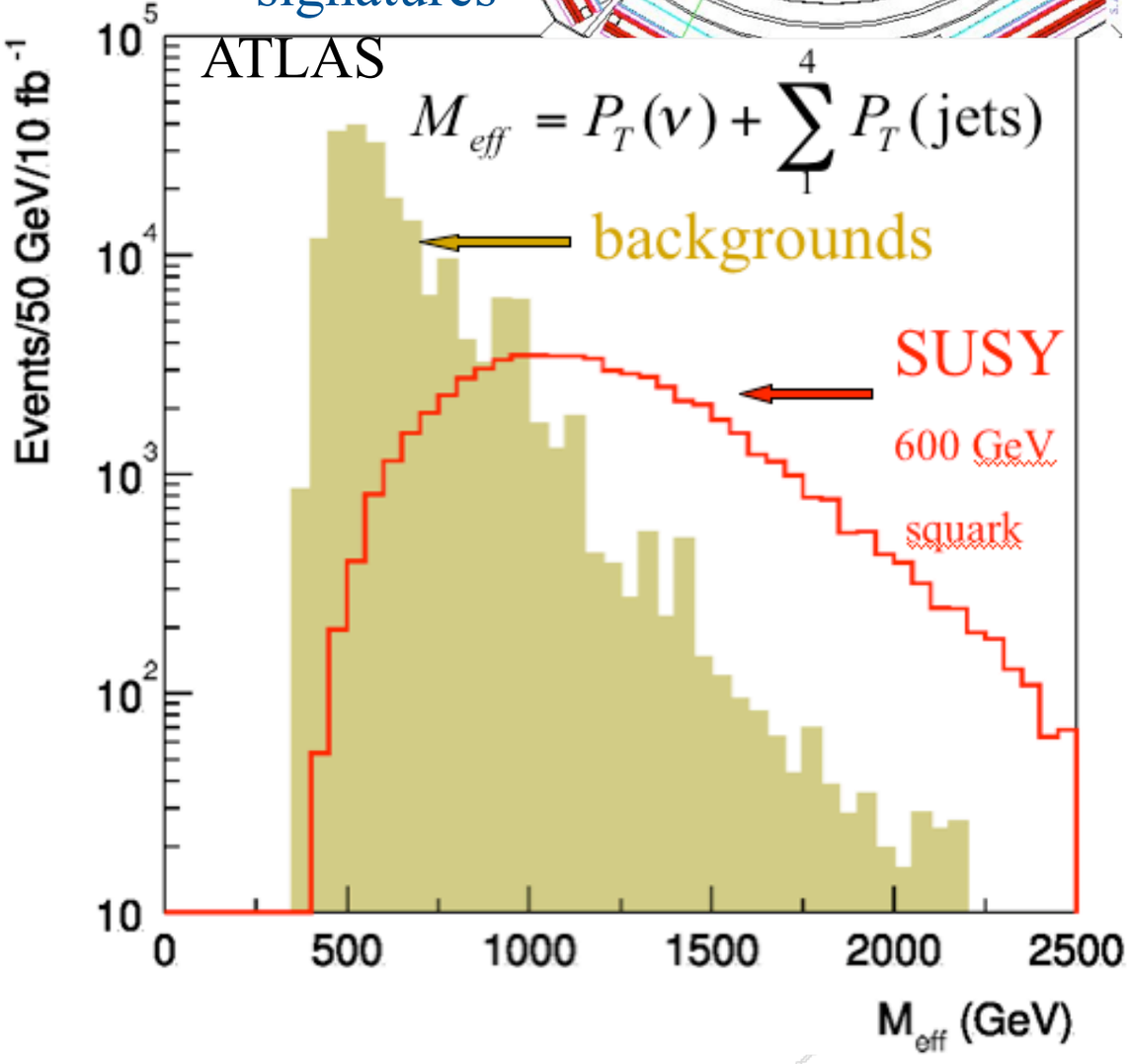
10 fb^{-1} expected in ~ first year

Full Higgs mass range accessible



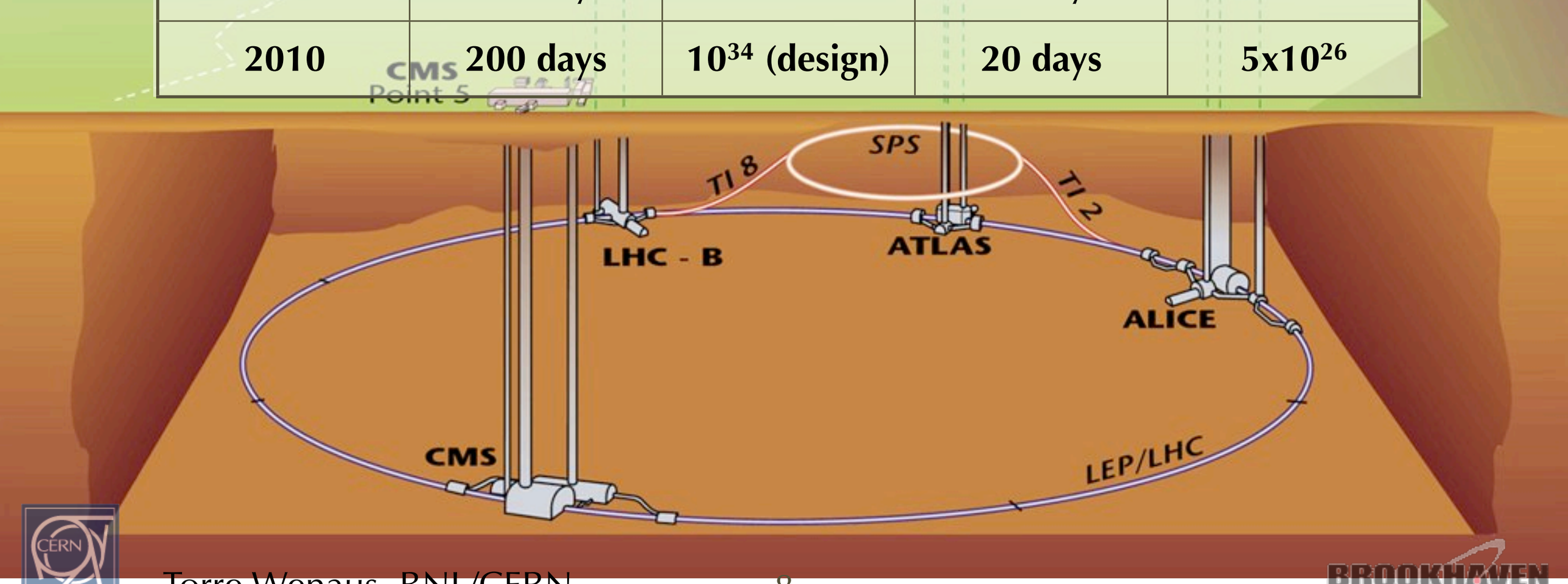
Dramatic SUSY signatures





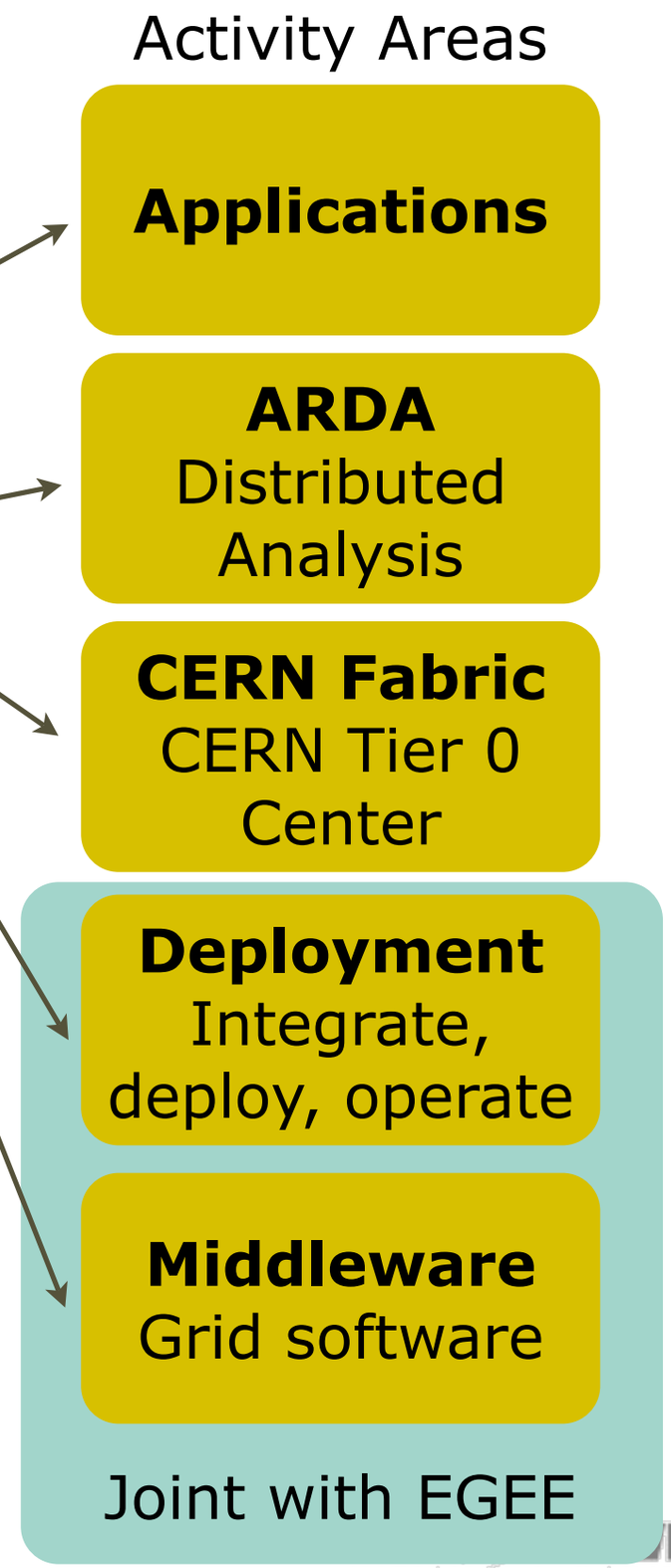
LHC Operation Overview

	pp beam time, L ($cm^{-2}s^{-1}$)		HI beam time, L ($cm^{-2}s^{-1}$)	
2007	**100 days**	**$5x10^{32}$**	**-**	**-**
2008	**200 days**	**$2x10^{33}$**	**20 days**	**$5x10^{26}$**
2009	**200 days**	**$2x10^{33}$**	**20 days**	**$5x10^{26}$**
2010	**200 days**	**10^{34} (design)**	**20 days**	**$5x10^{26}$**



LHC Computing Grid (LCG) Project

- Launched in March 2002 to **prepare, deploy and operate the computing environment for LHC data analysis**
  - Provide common physics applications software and development tools
  - Build and operate the LHC computing service
- ie. it is the LHC Computing Project
  - The grid is a tool towards achieving the goals
- Close collaboration between project, CERN, experiments, external sites, grid projects
- **Just completed: Experiment computing model documents**



Computing Models Cover

- **Data model**
  - Output stages, formats, sizes, rates, distribution
- **Analysis model**
  - Workflow, streams, (re)processing, data movement, interactivity
- **Distributed deployment strategy**
  - Computing tier roles, data management & processing across the tiers
- **Specifications for**
  - *Capacity* (processors, storage, network etc.)
  - *Capability* (middleware and other services)

Output Data Stages

ATLAS 1yr sizes 10^7s, 2x10^9event	
RAW	3.2PB
ESD	1PB
AOD	180TB
TAG	2TB

- SIMU - Simulated data
  - RAW format plus simulation info for debugging
- RAW - From detector DAQ
  - Reconstruction input; detector & software diagnostics, optimization
- ESD - Event Summary Data (Reconstruction output)
  - Reconstructed objects, constituents (allowing refitting, recalibration)
  - Several streams (express, calibration, [physics selections], ...)
  - Used for early/detailed analysis; most analysis based on AOD
- AOD - Analysis Object Data
  - Reconstructed objects for analysis, filtered and reduced from ESD
  - Many physics-selected streams
- TAG - Compact event summary with reference to event data

Data Sizes

All are approximate estimates
Trigger rates ~independent of luminosity

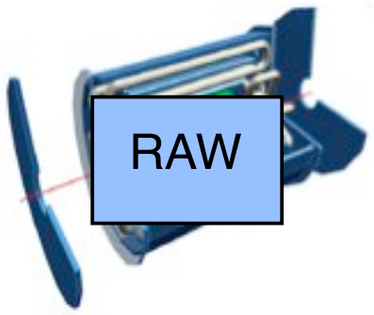
pp running

	Simu	Simu ESD	Raw	Trigger	Raw rate	Reco	AOD	Tag
ALICE	400kB	40kB	1MB	100Hz	100MB/s	200kB	50kB	10kB
ATLAS	2MB	500kB	1.6MB	200Hz	320MB/s	500kB	100kB	1kB
CMS	2MB	400kB	1.5MB	150Hz	225MB/s	250kB	50kB	10kB
LHCb		400kB	25kB	2kHz	50MB/s	75kB	25kB	1kB

Heavy ion running

	Simu	Simu ESD	Raw	Trigger	Raw rate	Reco	AOD	Tag
ALICE	300MB	2.1MB	13MB	100Hz	1.25GB/s	2.5MB	250kB	10kB
ATLAS			5MB	50Hz	250MB/s			
CMS			7MB	50Hz	350MB/s	1MB	200kB	

Analysis Model Overview



Experiment-wide activity
All events processed

Reprocessing
~3 times/yr
New calibrations, refinements in software, detector understanding

In early running, more frequent reprocessing and more use of RAW, ESD

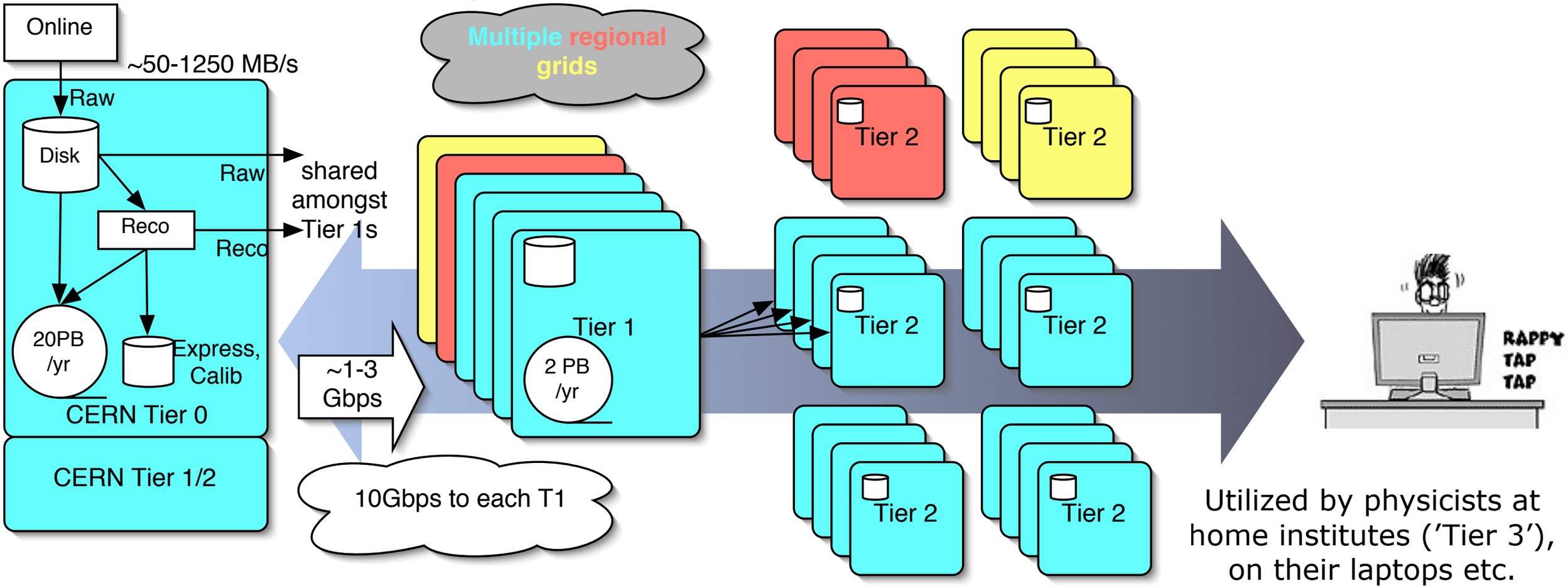
Physics group activity
O(20) groups, each selecting fraction of events

New selection pass
~monthly
Refinements in selections, software, physics understanding

Individual analysis

Constant 'chaotic' activity
Refining physics cuts and analysis software, MC comparison

Realizing the Models: The System



Utilized by physicists at home institutes ('Tier 3'), on their laptops etc.

CERN

Buffering from online
Express, calibration reco
Prompt first pass reco (not HI)
Archival copy of RAW, ESD data
Second copy sent to T1s
Reprocessing during shutdowns
Analysis facility with AOD, ESD
Analysis: ~100 users/expt

Tier 1s: ~6-10 per experiment, 10 total

Data import, serving, export
Share of RAW+ESD
Archive for locally produced data
ESD fraction, AOD on disk
Re-reconstruction
(Primary reconstruction if needed)
Calibration
Simulation
Analysis skims, user analysis

Tier 2s: ~3-5 per Tier 1, ~40 overall

Skims from Tier 1
ESD, AOD subsets on disk
User analysis - regional and/or by physics topic
Simulation (most experiments)
~20 active users (widely variable)

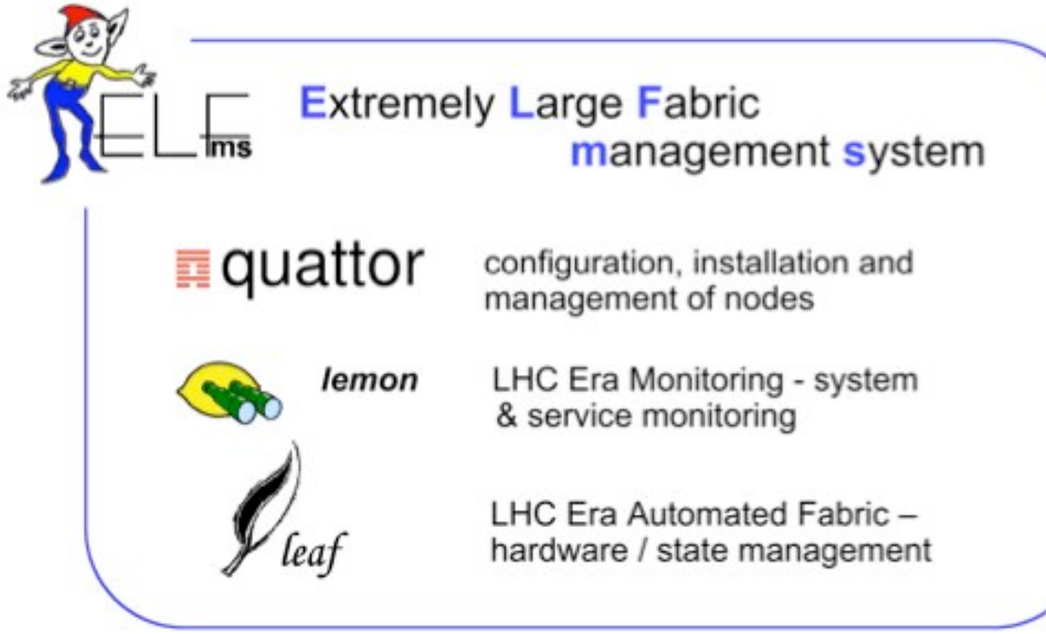
PC today ~1 kSi2k

Approximate total capacities circa 2008/2009

5 PB disk, 20 PB mass store, 20MSi2k	20 PB disk, 20 PB mass store, 45MSi2k	12 PB disk, 5 PB mass store, 40MSi2k

Realizing the System: CERN Fabric

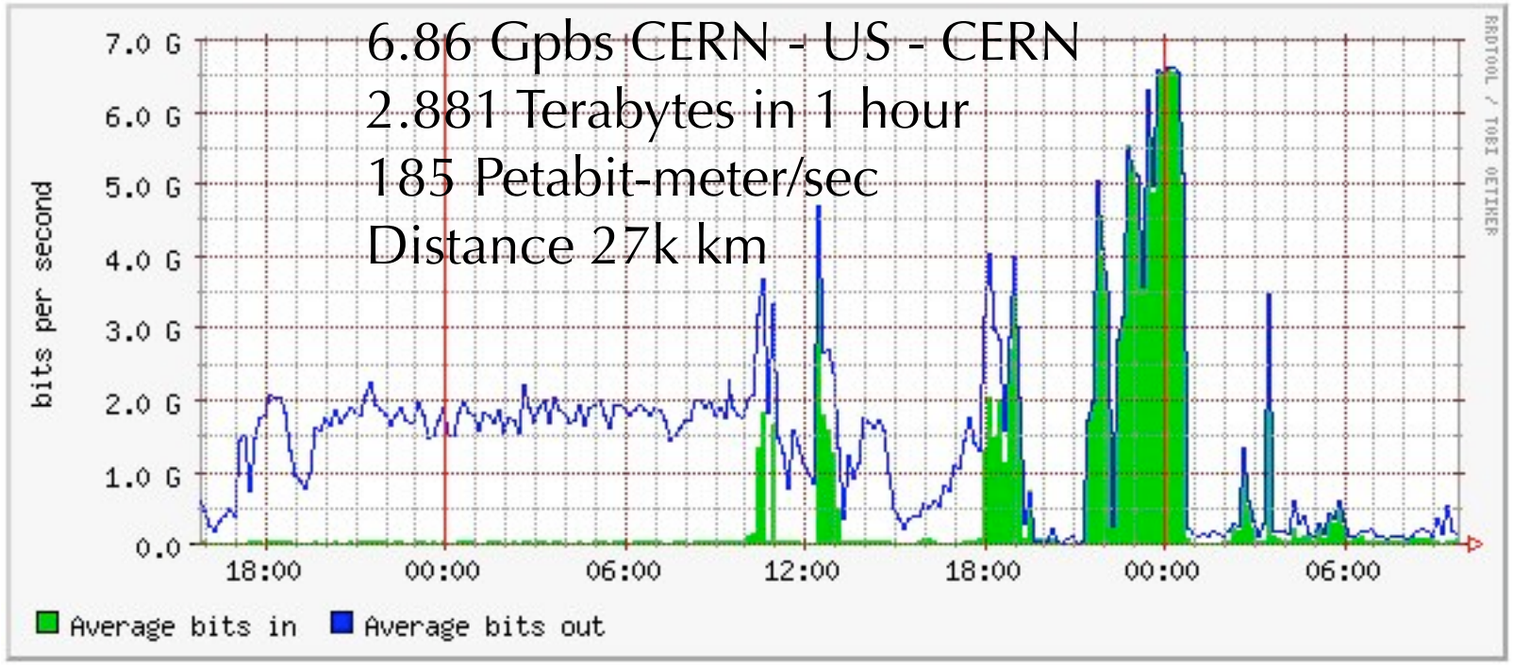
- Linux farm (production currently 32bit Intel)
  - "Moore's Law" evolution being watched, not a major worry (shift from GHz to multicore focu helps power/heat issues)
- Automation well developed for installation, configuration, management, monitoring
- Standard 'Scientific Linux' distribution in collaboration with Fermilab, Red Hat
  - Addresses OS uniformity, licensing
- CASTOR mass storage system
  - Recent scaling problems expected to be resolved by new release
- Linux-based Oracle physics DBs
- CERN physical plant - power (2.5MW), space (computing center expansion)
- **On track for 7000 boxes in 2008**



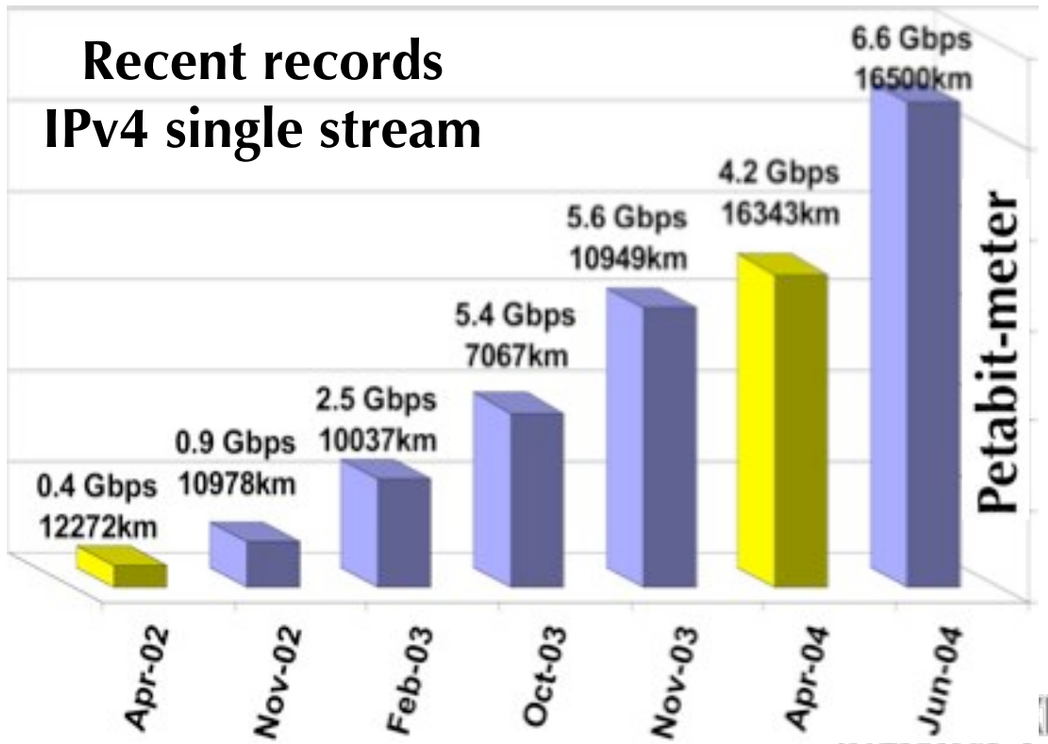
Networking

- Bandwidth growth on track: 10 Gbps to most T1s in 2005
  - Growth >> Moore's Law
- Affordable due to overbuilt fibre and commodity hardware
  - But who pays is an issue
- Growing 'Digital Divide' to less privileged regions is a problem
  - HEP proactive in addressing this (cf. H. Newman talk, CHEP04)
- Focus is now on reliable TB scale transfers

Internet land speed record, Caltech/CERN et al, Nov '04

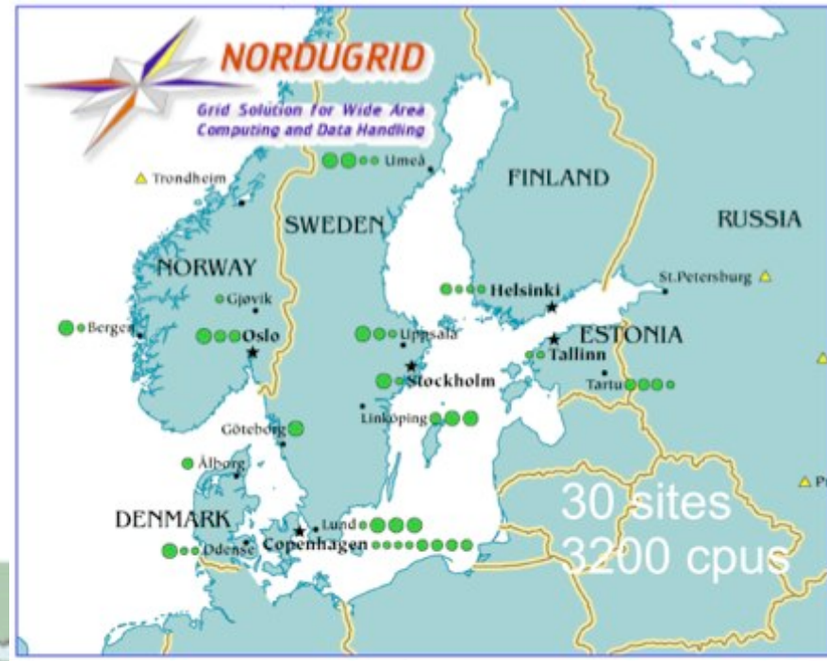


IPv4 multi-stream, Linux, FAST TCP (Caltech)



Realizing the System: Deployment

Three distinct (but collaborating) regional grids
Three distinct (but overlapping) middleware 'flavors'
(Multiplicity of grids is a disappointment but not a surprise)



Goal: general service for all experiments, operated for them by regional centers

LCG Project
- operates LCG/EGEE grid
- coordinates exploitation of wider resources

LCG/EGEE

Middleware foundation is US-developed Virtual Data Toolkit

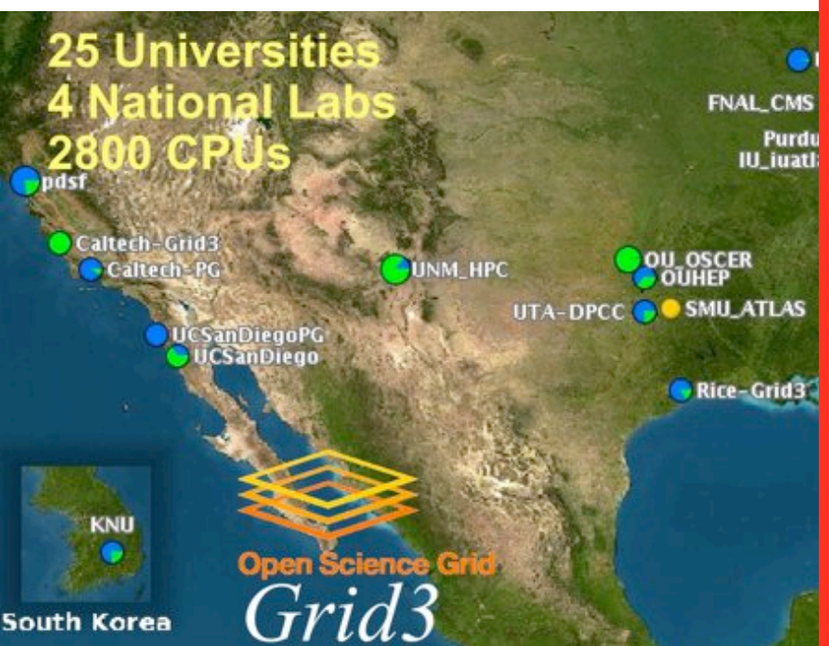
(Re)engineering higher levels (gLite) based on previous generation deployment experience (EDG, LCG-1, LCG-2)

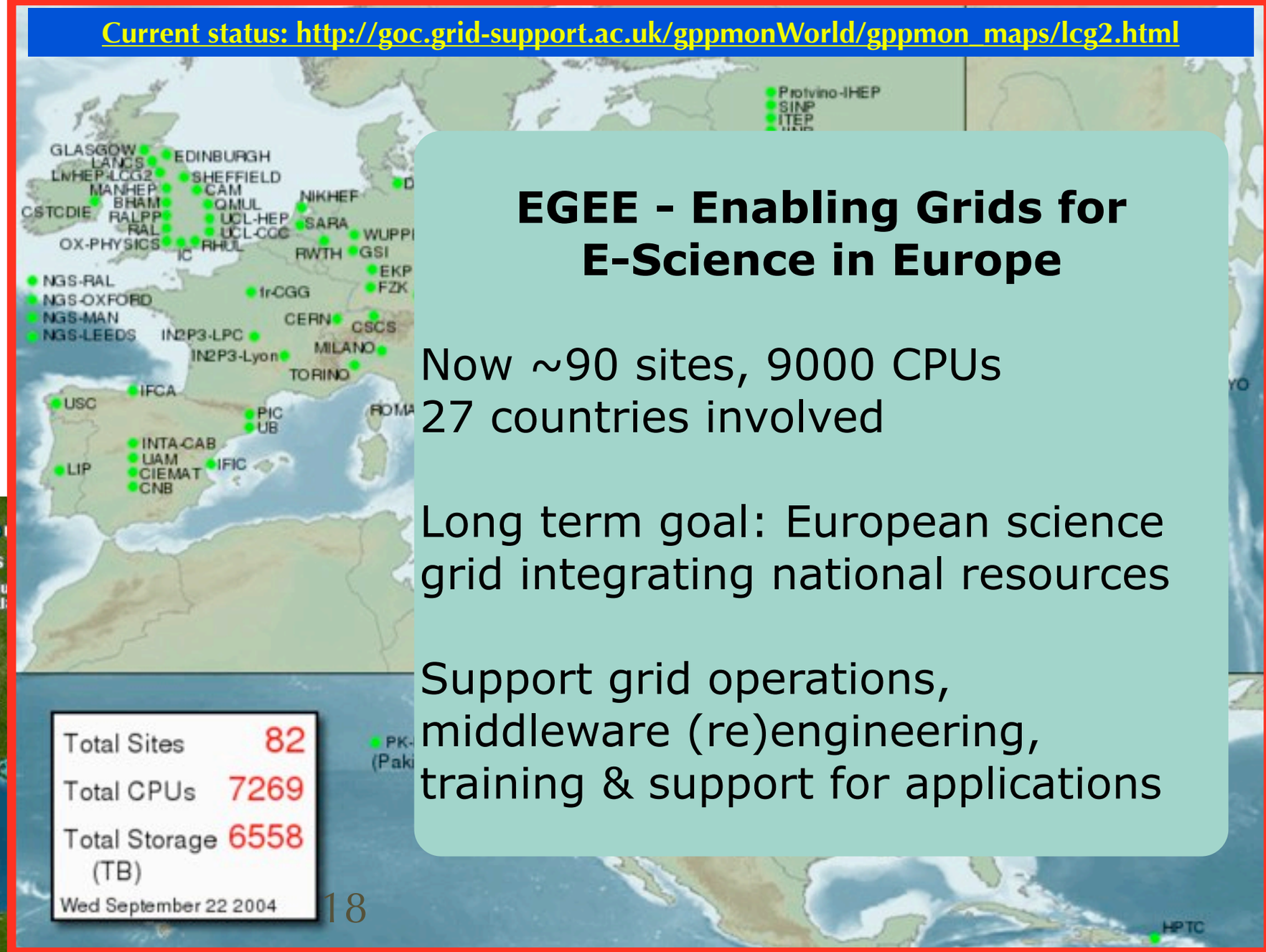
No time for NorduGrid slide, apologies!

Transition in 2005 from LCG-2 (production focus) to gLite (adding analysis)

gLite Goals:

- **leverage existing work, e.g. ALICE's AliEn**
- **short development cycles**
- **deploy & validate early in distributed analysis**





EGEE - Enabling Grids for E-Science in Europe

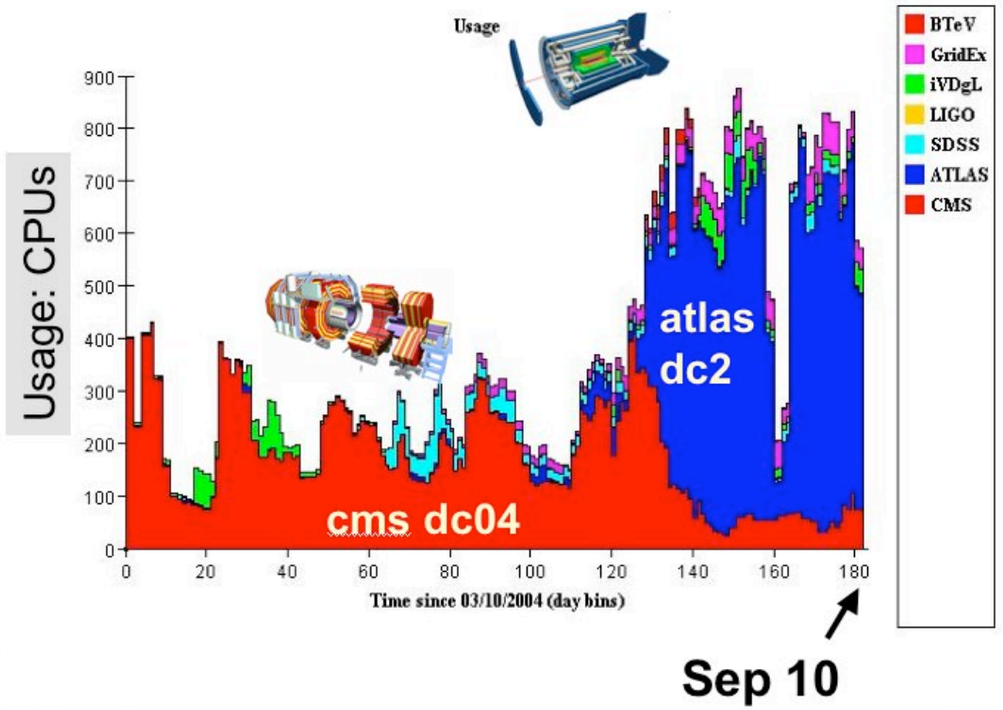
Now ~90 sites, 9000 CPUs
27 countries involved

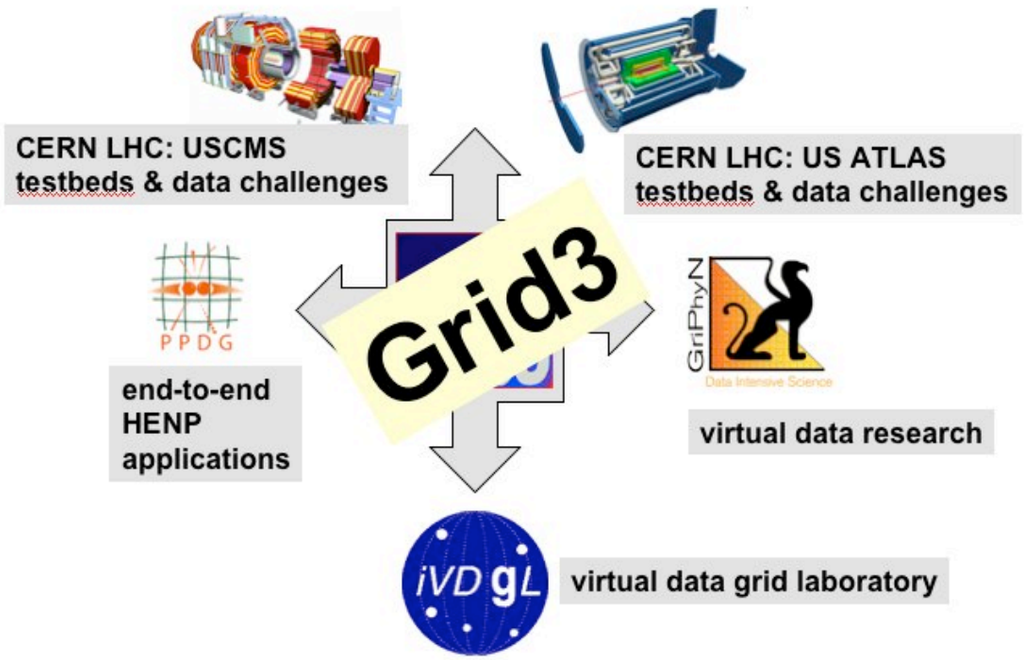
Long term goal: European science grid integrating national resources

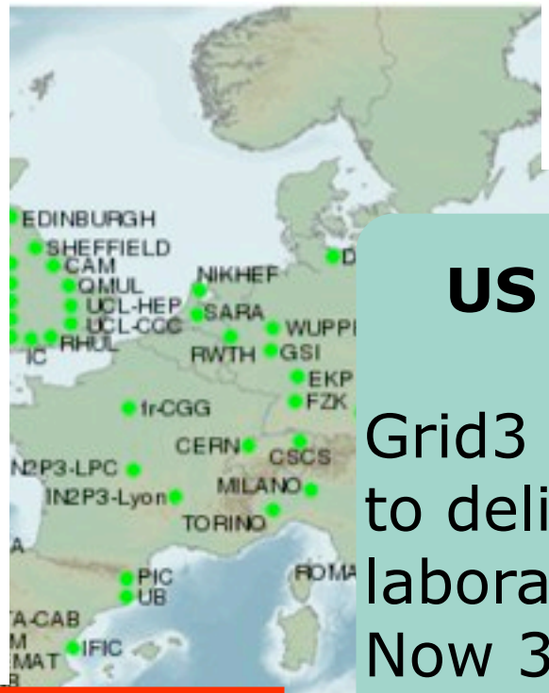
Support grid operations, middleware (re)engineering, training & support for applications

US Grid3/OSG

Middleware foundation is VDT (Globus, Condor, ...)
Supports multiple VOs, grid monitoring and operations
Higher level services provided by individual VOs





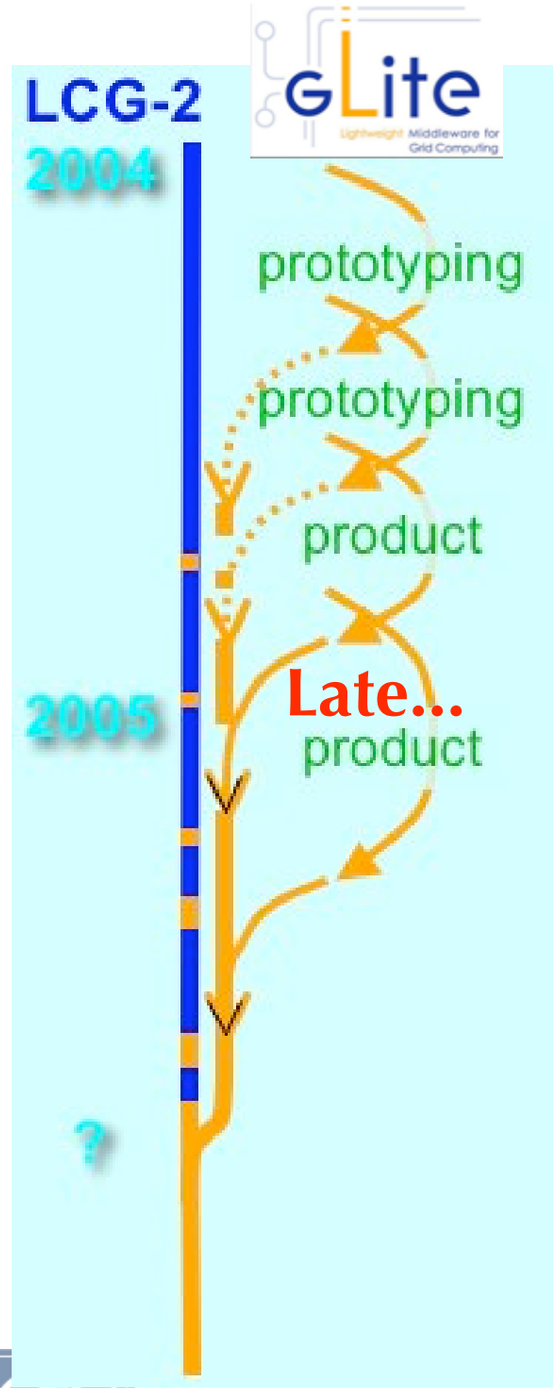


US Grid3/Open Science Grid

Grid3 tied together US grid projects to deliver an operating grid laboratory for e.g. LHC DCs.
Now 35 sites, 3500 CPUs

Now evolving into Open Science Grid (OSG) a US multi-science grid
- improve data management, authorization, monitoring services
- LCG Service Challenges
- Improve *grid interoperability*

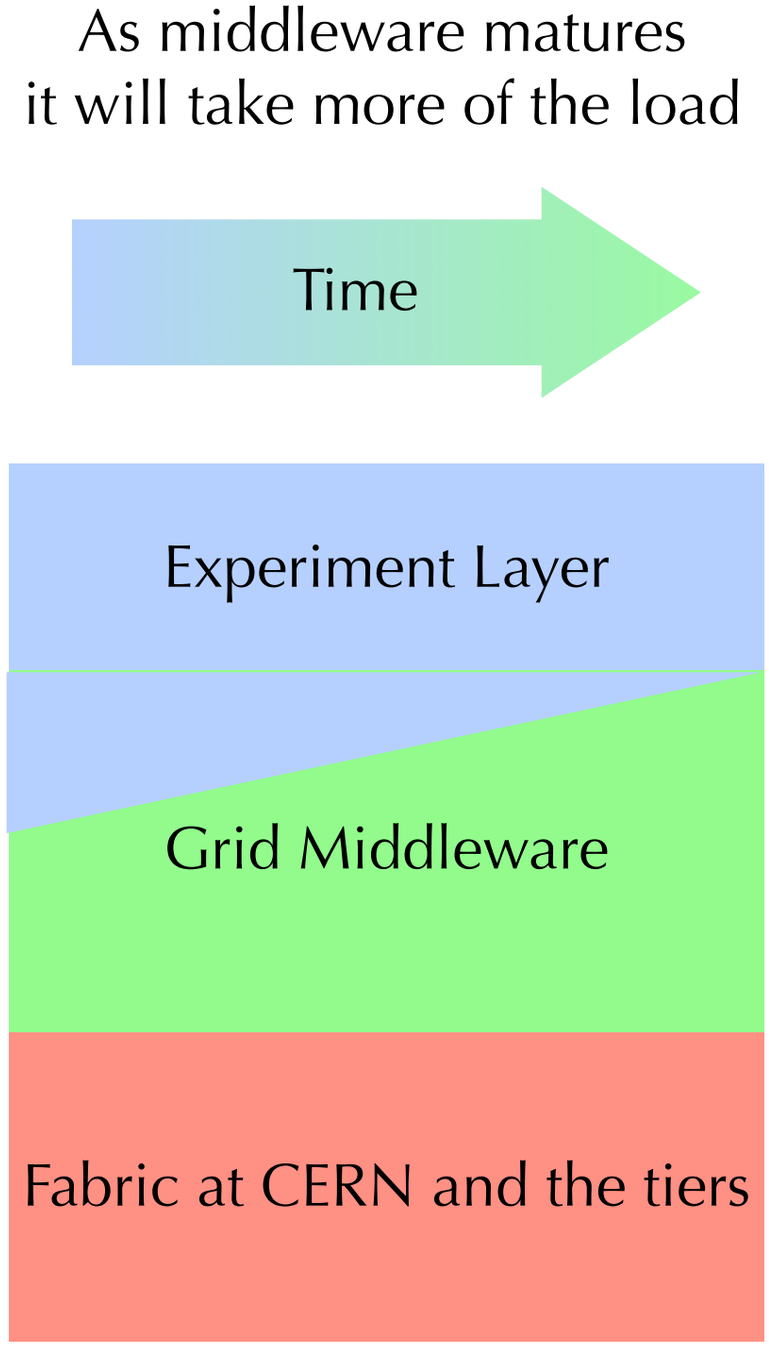
Realizing the System: Middleware



- Original ('02) LCG plan: "The LCG is not a middleware project"
  - Was to be delivered to us... but too little, too late
  - Feature set, performance, scalability disappointing
- New ('04) plan: Middleware "re-engineering" as part of the LCG program, in collaboration with EGEE
  - Current concern: avoid a repeat
  - New EGEE gLite suite is late, expectations are cautious
- Being addressed in several ways
  - Acceleration and monitoring of gLite development
  - Focus on the essential middleware needed for startup
  - Experiment caution: put minimal demands on middleware for startup -- a 'thicker experiment layer'
    - Incorporate more middleware as it matures -- as measured by quantitative performance metrics

Current Middleware Priorities

- Data management
  - We are *data intensive*; middleware presently weak in this area
  - Storage management, file transfer, catalogs
  - Experiment layers can and do help here, leveraging basic middleware services (grid ftp, storage interfaces,...)
    - Also examining and/or adopting solutions from current HEP generation
      - e.g. xrootd distributed data manager (BaBar)
- Virtual organization management
  - **One** standard for authentication/access: one certificate provides global identity, access rights
  - Managing multiple virtual organizations on shared facilities
  - This *must* come from the middleware



Middleware Collaboration

Dealing with opposing forces is a challenge!

HEP users	**Middleware providers**
Stability	Innovation
Mature, proven solutions	New/emerging technologies
Performance, scalability, reliability	Functionality
End-to-end systems view	Component view
Immediate & near term needs	Young technology
"Not invented here" syndrome	
Uniform global computing	Regional services
HEP requirements	Funder requirements
Secure but easy access	Security

Motivating Middleware Collaboration

But the collaboration has strong motivations

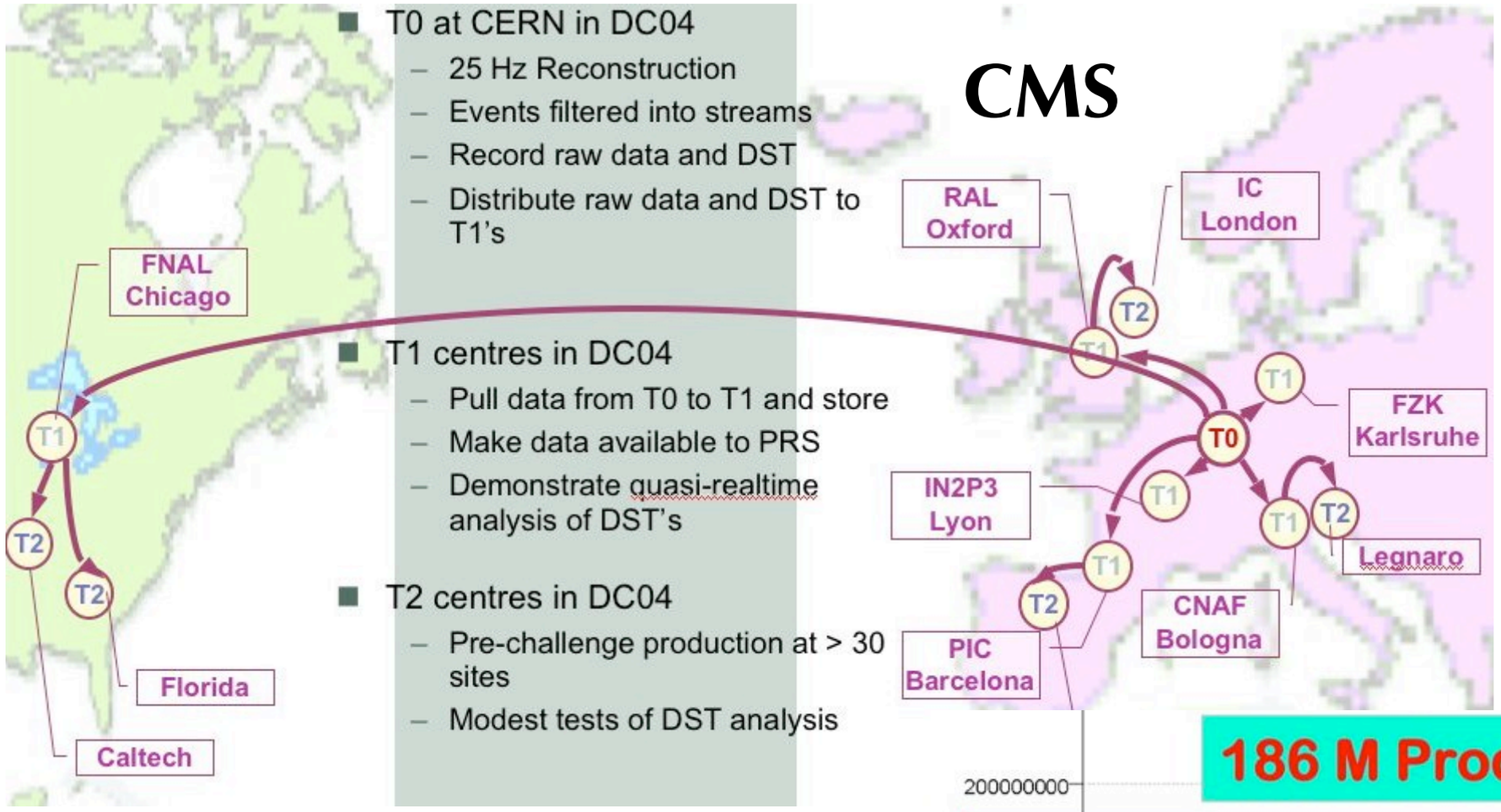
HEP users	**Middleware providers**
Tight computing funding	Well funded grid projects
Leading edge needs, presently unmet	Scientific apps indicate IT trends
Real applications, realistic scales	Need real-world deployers, feedback
Hands-on approach to collaboration	Eager for close, engaged collaboration
Better, earlier, non-redundant tools from interdisciplinary collaboration	
Agreement on services, interfaces now enables interworking later	
Tight computing funding	Well funded grid projects
Tight computing funding	Well funded grid projects
…	…

2004 Data Challenges

Distributed production at large scales by all four experiments

CMS

- T0 at CERN in DC04
  - 25 Hz Reconstruction
  - Events filtered into streams
  - Record raw data and DST
  - Distribute raw data and DST to T1's
- T1 centres in DC04
  - Pull data from T0 to T1 and store
  - Make data available to PRS
  - Demonstrate quasi-realtime analysis of DST's
- T2 centres in DC04
  - Pre-challenge production at > 30 sites
  - Modest tests of DST analysis



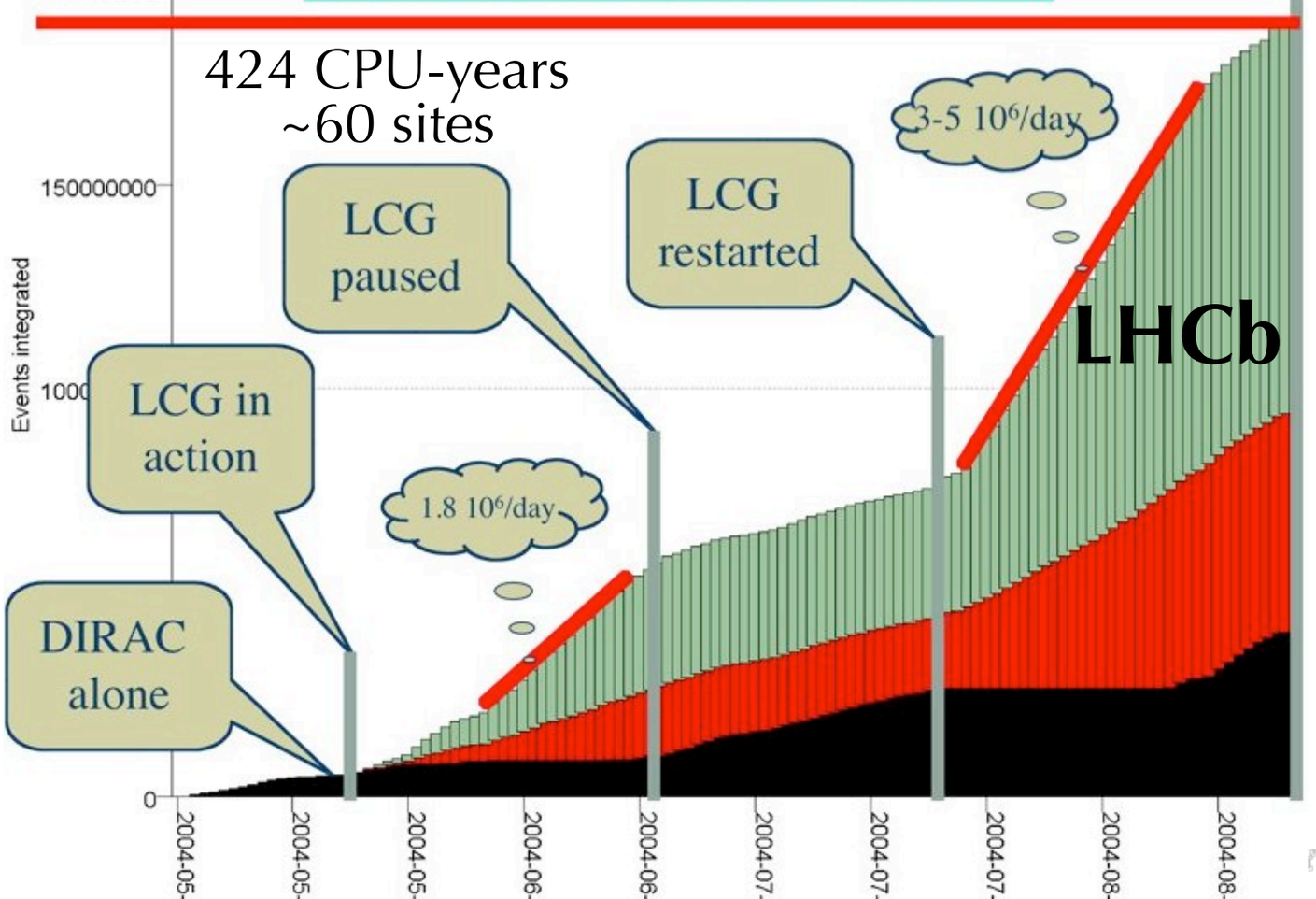
	Jobs(k)	%Sub	%Remain
Submitted	211	100.0%	
Cancelled	26	12.2%	
Remaining	185	87.8%	100.0%
Aborted (not Run)	37	17.6%	20.1%
Running	148	70.0%	79.7%
Aborted (Run)	34	16.2%	18.5%
Done	113	53.8%	61.2%
Retrieved	113	53.8%	61.2%

LHCb

LCG Efficiency: 61 %

186 M Produced Events

424 CPU-years
~60 sites



2004 Data Challenge Outcomes

- Exercised facilities, middleware, production systems, experiment software at large scales
- **Massive production on LCG, Grid3, Nordugrid worked**
  - Successful shared, opportunistic resource usage
- Many old and familiar problems: full/dead disks, reboots, …
  - Dealing with them still takes substantial effort in the 'grid era'
- Areas needing improvement identified
  - Low grid production efficiency; ~60% characteristic
  - Operational stability; robustness against failures (network, unexpected shutdowns, …)
  - Monitoring, logging, difficulty of troubleshooting
  - Packaging, installation, configuration tools & procedures
  - Workload and data management systems
  - ***Scalability of implementations: data management (file cataloging and access), job submission, information system***
- **Analysis not a focus in '04 DCs: this is coming in 2005/2006**

Common Applications - LCG Applications Area

- A common effort on physics applications software development
- Motivated by (common needs and) inadequate resources to support independent efforts
- Key mandates to deliver missing pieces of the required software and infrastructure
  - Provide a physics data storage framework
  - Guide the completion and physics validation of detector simulation tools
  - Provide software for a component based C++ infrastructure integrating with experiment frameworks and the ROOT analysis system

AA Projects

SEAL
Core libraries, services

Persistency
Physics data storage (POOL)

Simulation
Validation, tools, services

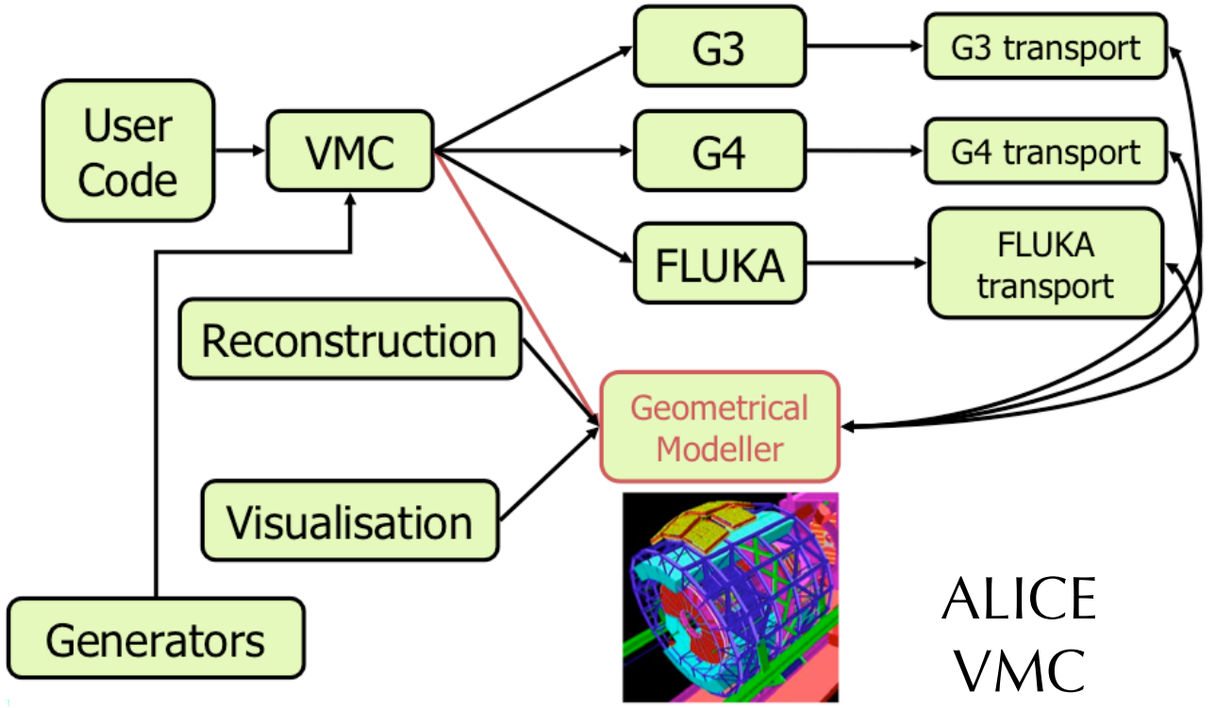
PI
Physicist Interfaces

SPI
Development, distribution support

Physics Data Storage

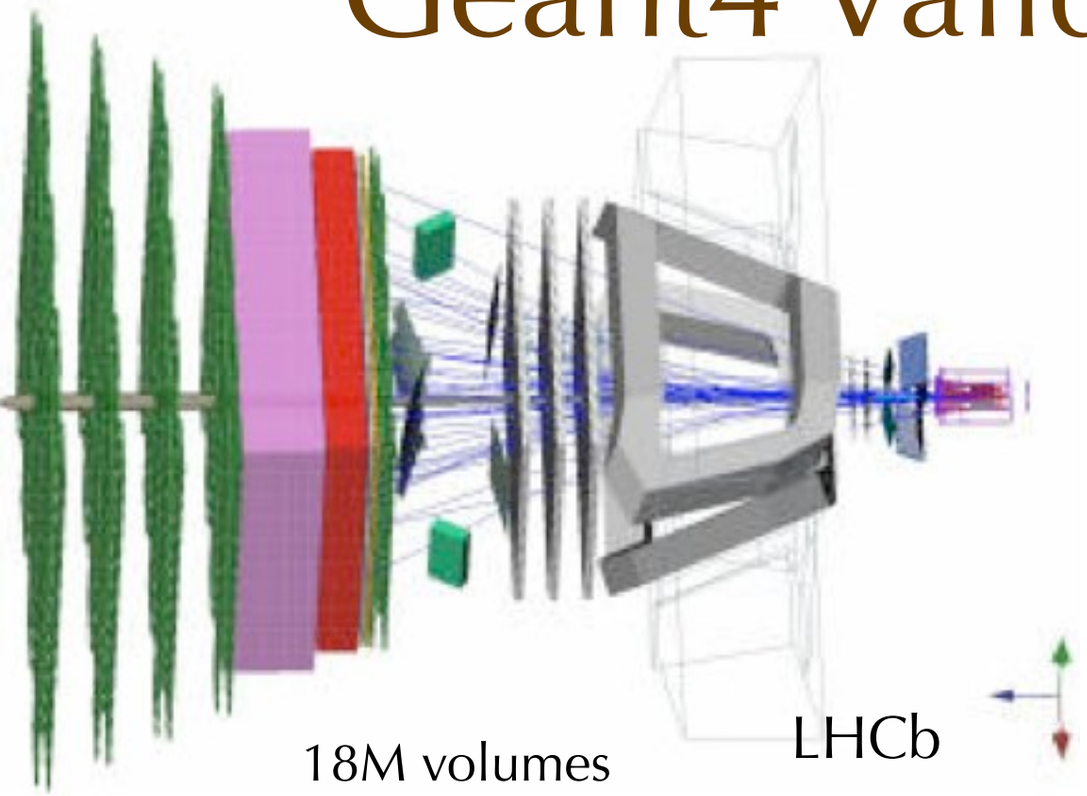
- A new persistency system, POOL, provides large scale data storage for the C++ based software systems of the LHC experiments
  - ROOT I/O for bulk object-based data storage
  - Complemented by relational DB technology for cataloging and data lookup
- Validated in 2004 Data Challenge deployments by the three experiments using it (ATLAS, CMS, LHCb); ~400TB stored
  - ALICE uses ROOT I/O directly
- A critical issue 3 years ago, now resolved
- Current development focus: conditions data (calibration, alignment,...)
  - POOL extended to provide relational database support
  - New conditions database system to be released in March
- Current priority issues:
  - Schema evolution strategies -- Tools in place, how do we use them
  - *Distributed, scalable* access to data -- Addressed in new distributed database deployment project, strongly leveraging Run 2 experience, tools

Simulation

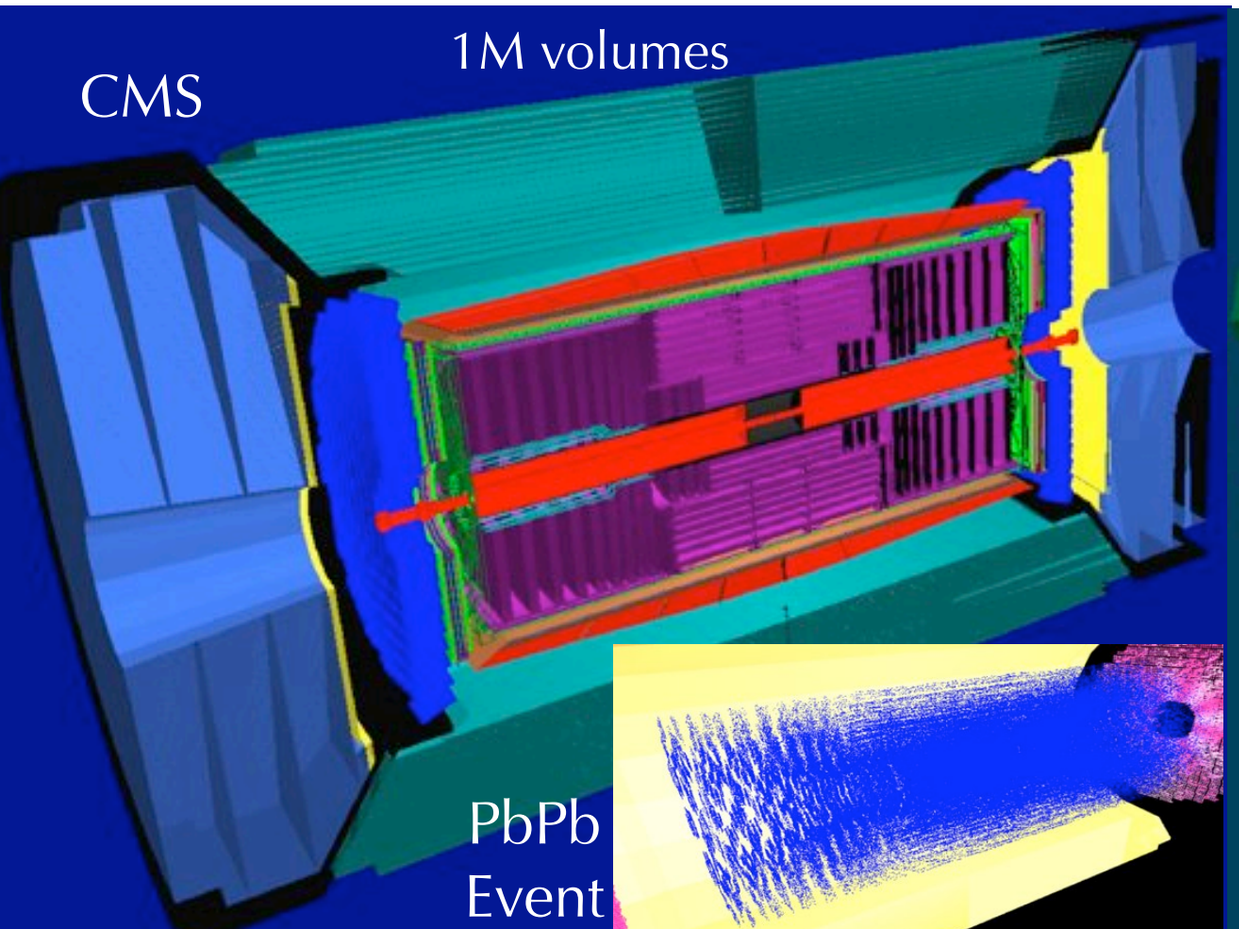


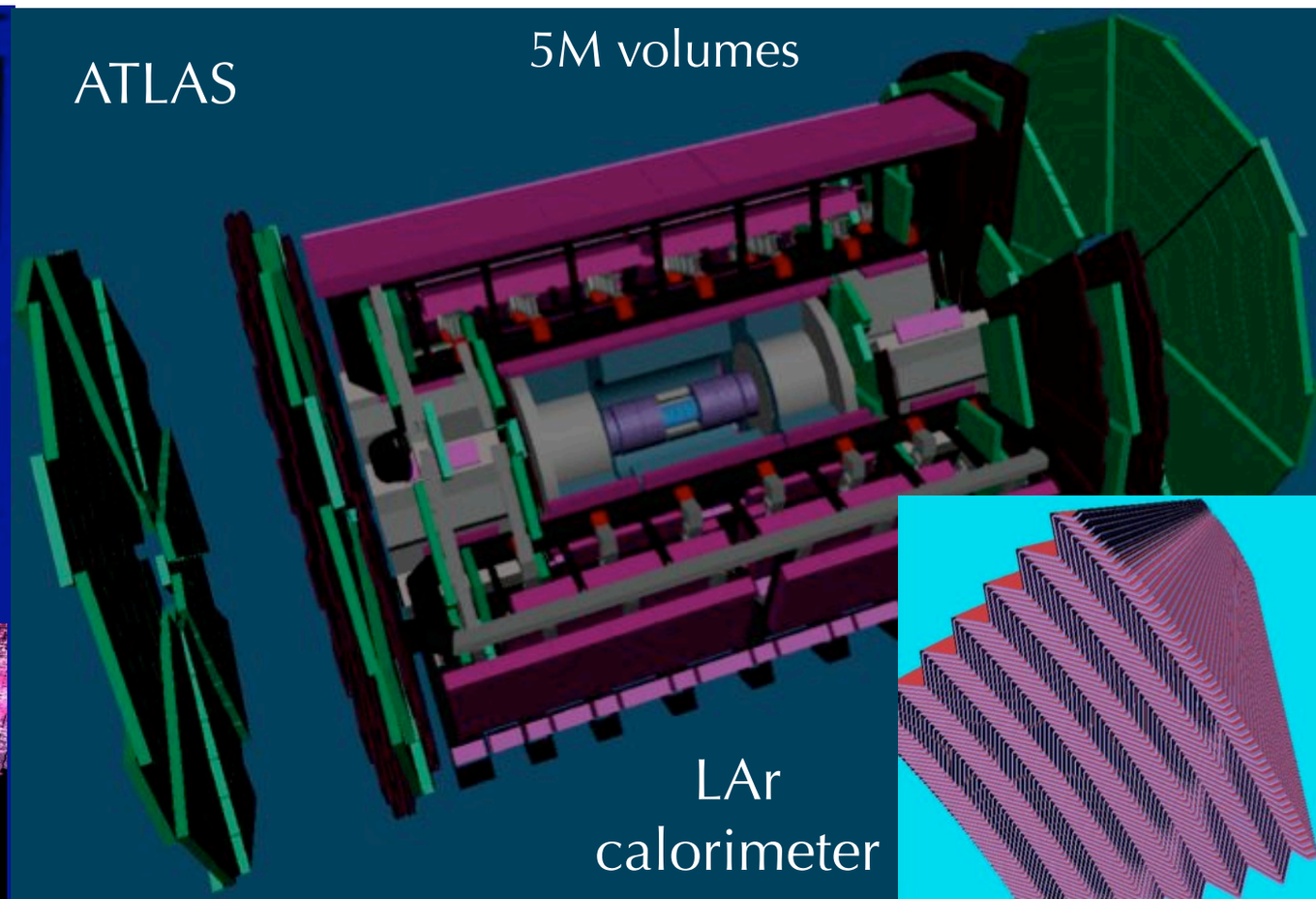
- Robust and validated simulations are in place for all experiments
  - A source of worry in the past
  - ATLAS, CMS, LHCb: Based on Geant4 toolkit
  - ALICE: Virtual Monte Carlo, supporting multiple simulation engines (Geant3, FLUKA, Geant4) in a single simulation environment
    - An ALICE/ROOT effort, not LCG/AA
  - Demonstrated in large scale simulation production in 2004 DCs
  - Physics of Geant4 and FLUKA validated against data, e.g.
    - modelling microscopic collisions in thin materials for tracker occupancy
    - e/π ratio (to few %) governing non-linearity in hadron calorimeter response for jet energy measurement

Geant4 Validation in 2004 DCs



- Robust: crash rates of order 1 per 1M events
- Highly detailed detector models, physics processes greatly refined since Geant3
- 20-30′ for TeV-scale dijet event (ATLAS, CMS); 180′ for CMS PbPb event (55k tracks); program size 200-400MB





ROOT and SEAL

- ROOT is HEP's most popular analysis tool (in C++ at least!)
  - Plays central roles in all experiments' software and the LCG, e.g.
    - ROOT I/O for POOL data storage
    - Used for analysis in all experiments
    - Directly used as basis for framework by ALICE
- SEAL created to provide framework-level tools for experiments not using ROOT directly
  - i.e. its existence reflects different approaches in the experiments
- Since SEAL was established, ROOT/SEAL convergence has progressed greatly and is still developing
  - Greater penetration of ROOT usage in all experiments
  - SEAL efforts seen as useful by ROOT as well as experiments
    - SEAL-developed ROOT-Python integration taken up by ROOT

Ramping up: Service Challenges

SC2: Robust data transfer with 6 sites, >1 reaching sustained 500 MB/s with CERN

SC3 Challenge: SC2 sites at 50% scale. 60 MB/s to tape @ each T1 from CERN disk

SC3 Service: SC2 + new sites. Stable service for computing model tests at 50% scale

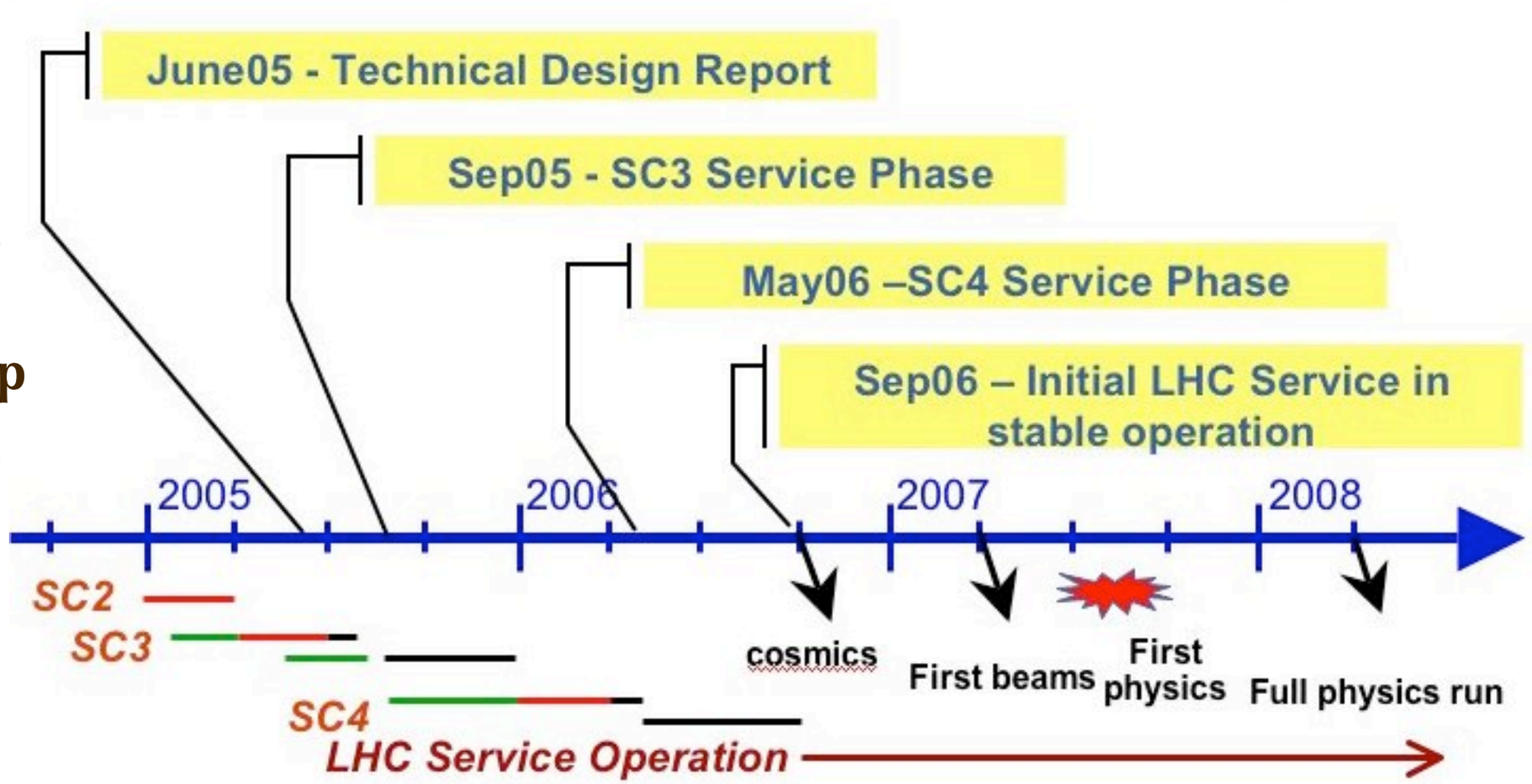
SC4 Preparation: Boost CERN production to 3.6 GB/s, T1s to full nominal rate

SC4 Challenge: T0, all T1s, major T2s operating at full target rates (~2 GB/s at T0)

100% computing model validation, all tiers & experiments

SC4 Service becomes permanent LHC service. Capacity milestone 3 mo before physics

Newly organized to prepare and validate all tiers and the middleware for scaled up DCs and for datataking



Conclusion

- Computing is (of course!) on the critical path for all
- Experiment software, fabrics, backbone networks generally on track
- The infrastructure to bind them and drive analysis at scale is less clear
  - Data intensive distributed computing is, as expected, hard
- Reflected in present priorities:
  - Building the experiment data management systems
  - Scalable, distributed access to database-resident data
  - Extending distributed processing from production to support for end users
  - Ramping end-user distributed analysis, for its own sake and to start exercising and understanding 'chaotic' grid usage by individuals
  - Scalable and robust middleware
- Pragmatism is the word of the day
  - Analysis readiness is foremost: follow the means that best gets us there
  - Ingredients from the current HEP generation, LHC experiment developments, and leading edge tools from the grid community are all in the mix

For More Information

- LHC experiment computing model documents (January 2005)
  - http://lcg.web.cern.ch/LCG/PEB/LHCC/expt%5Freqts/
- Comprehensive recent snapshot of LHC and other HEP computing: CHEP 2004 conference (September 2004)
  - http://chep2004.web.cern.ch/chep2004/

Supplemental

Tier 1 Centers As Of Jan 2005

				ALICE	ATLAS	CMS	LHCb	
1	GridKa	Karlsruhe	Germany	X	X	X	X	4
2	CCIN2P3	Lyon	France	X	X	X	X	4
3	CNAF	Bologna	Italy	X	X	X	X	4
4	NIKHEF/SARA	Amsterdam	Netherlands	X	X		X	3
5	Nordic	Distributed	Dk, No, Fi, Se	X	X			1
6	PIC	Barcelona	Spain		X	X	X	3
7	RAL	Didcot	UK	X	X	X	X	4
8	Triumf	Vancouver	Canada		X			1
9	BNL	Brookhaven	US		X			1
10	FNAL	Batavia, Ill.	US			X		1
11	ASCC	Taipei	Taiwan		X	X		2
				6	10	7	6	

x – announced at January GDB

Blue: LCG grid

J. Shiers, from comp model documents

Total Resource Estimates

Estimated 2008 Resource requirements

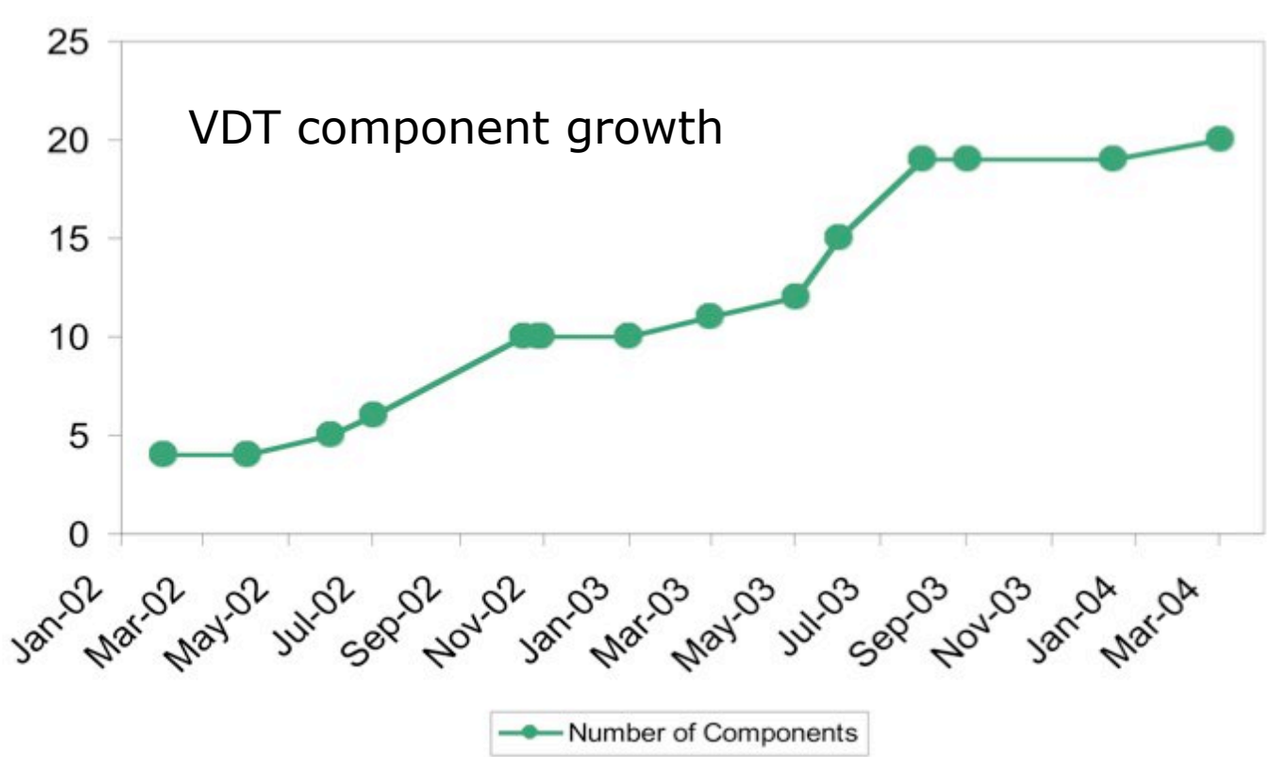
	CPU MSi2k	Disk PB	Tape PB
ALICE	31	8.3	10.3
ATLAS	59	31	16
LHCb	13	3.3	3.4
CMS	40	13.7	16.6

Virtual Data Toolkit (VDT)

A success in standardized middleware

VDT or components used in all grids

Not without problems of slow maturation of feature sets, performance, scalability



VDT Providers, Components

- **Condor Group:** Condor job submission system, Fault tolerant shell, ClassAds
- **Globus Alliance:** Job submission (GRAM), Data transfer (GridFTP), Information service (MDS), Replica location service (RLS)
- **EDG, LCG:** Make Gridmap, Certificate management, Glue schema services
- **ISI, UC:** Chimera, Pegasus
- **NCSA:** MyProxy, GSI OpenSSH, UberFTP
- **LBL:** PyGlobus, NetLogger
- **Caltech:** MonaLisa grid monitor
- **VDT:** VDT system profiler, configuration
- **Others:** KX509, DRM 1.2, Java, FBSng job manager

Managed by U Wisconsin

Service Challenges

- Newly organized to follow on from 2004 DCs
- Prepare and validate all tiers and the middleware for scaled up Data Challenges and ultimately for startup
- Supporting the sustained average rates below requires capability to
  - run for extended periods at double these rates
  - sustain 2-4 times these rates on the network

Nominal raw + ESD data rates out of CERN and onto tape at a Tier 1

	ALICE	ATLAS	CMS	LHCb	Total
CERN	120-600 MB/s	750 MB/s	700 MB/s	100 MB/s	1.7-2.2 GB/s
Average T1	20-100 MB/s	75 MB/s	100 MB/s	60 MB/s	260-340 MB/s

Connectivity Timeline

	CERN	Tier 1s
Sep 2005	5-10 Gbps	1-10 Gbps
Jan 2006	35 Gbps	10 Gbps at >5
July 2006	70 Gbps	10 Gbps